Comprehensive Analysis of Fraudulent Transactions

Objective

The primary goal of this project was to identify patterns, trends, and anomalies in fraudulent transactions using advanced exploratory data analysis (EDA). We aimed to uncover actionable insights by analyzing transaction amounts, times, demographics, and other key features.

Dataset Overview

Key Features

- **Amount**: The transaction value.
- **Class**: Indicates whether a transaction is fraudulent (`1`) or non-fraudulent (`0`).
- **Hour**: The hour at which the transaction occurred.
- **Transaction_Channel**: Online or in-store transaction.
- **Transaction_Location**: Local or international transaction.
- **Merchant_Type**: The category of the merchant.
- **Cardholder_Age**: The age of the cardholder.
- **Transaction_History**: Historical transaction amounts of the cardholder.

Summary of Columns

- Number of Rows: `X`
- Number of Columns: `Y`

Analysis Workflow

1. Identifying Transaction Patterns

Key Question: How can we identify patterns in fraudulent transactions based on transaction amount, time of day, and merchant type?

Steps Taken:

- **Pandas Analysis**: Descriptive statistics were calculated for fraudulent and non-fraudulent transactions, grouped by hour and merchant type.
- **NumPy Metrics**: Percentiles, means, and other statistical measures highlighted specific trends in fraud.
- **Visualizations**: Matplotlib and Seaborn visualizations (boxplots, barplots) showed fraud trends across hours and merchant types.

Insights:

- Fraudulent transactions are dominated by **small amounts** (median: `$34.22`) and occur most frequently during **late-night hours** (22:00-23:00).
- Fraud is most prevalent in the **grocery** and **health** merchant categories.

---

2. Vulnerability by Transaction Channel and Location

Key Question: Which transaction channels (online/in-store) and locations (local/international) are more prone to fraud?

Steps Taken:

- **Pandas Analysis**: Transactions were grouped by channel and location, then analyzed for fraud rates.
- **NumPy Metrics**: Fraud rates were calculated for online vs. in-store and local vs. international transactions.
- **Visualizations**: Stacked barplots highlighted fraud distribution by channel and location.

Insights:

- **Online transactions** are far more vulnerable (fraud rate: `16.33%`) than in-store (`4.70%`).
- Fraudulent activity is concentrated in **international transactions** (`14.71% fraud rate`) compared to local (`6.78% fraud rate`).

---

3. Fraudulent Demographics

Key Question: Are younger cardholders more likely to be involved in fraudulent transactions compared to older cardholders?

Steps Taken:

- Age groups were created using Pandas, and transactions were grouped by age and class to calculate fraud rates.
- NumPy metrics like mean and median ages provided a deeper understanding of fraud demographics.
- Visualizations (barplots and line plots) revealed fraud patterns across age brackets.

Insights:

- Fraud is overwhelmingly concentrated in the **16–25 age bracket** (fraud rate: `99.24%`).
- Older age groups (41–60 and 61+) showed no fraudulent activity.

---

4. Transaction History and Fraud

Key Question: How does transaction history differ between fraud and non-fraud cases?

Steps Taken:

- Descriptive statistics were computed for transaction histories grouped by class.
- NumPy percentiles and mean differences quantified fraud vs. non-fraud disparities.
- Visualizations (boxplots, histograms) compared transaction histories.

Insights:

- Fraudulent transactions have significantly lower historical transaction amounts (mean: `$75.37`, median: `$75.52`) compared to non-fraudulent ones (mean: `$300.25`).
- Fraudulent transaction histories are tightly clustered within a narrow range.

---

5. Outliers and Anomalies

Key Question: Can we identify outliers or anomalies indicative of potential fraud based on transaction amount, time, or other features?

Steps Taken:

- IQR was used to detect outliers in transaction amounts, separated by fraud and non-fraud cases.
- Time-based patterns were analyzed for anomalous hours with high fraud activity.
- Scatter plots and boxplots visualized these outliers.

Insights:

- Fraudulent transactions show minimal outliers, reflecting consistent behavior.

- Non-fraudulent transactions exhibited greater variability, with `6.96%` flagged as outliers.

---

6. Machine Learning Model Implementation

Objective

To build a predictive model to detect fraudulent transactions based on features such as transaction amount, channel, location, and merchant type.

Workflow

1. **Data Preprocessing**:
    - **Categorical Features**: Encoded using one-hot encoding (e.g., `Transaction_Channel` converted into binary columns for "Online" and "In-Store").
    - **Numerical Features**: Standardized using scaling (e.g., `Amount` and `Transaction_History` scaled to have mean 0 and standard deviation 1).
    - Preprocessing was automated using a Scikit-learn `Pipeline` for efficiency and reproducibility.
2. **Train-Test Split**:
    - Dataset was split into **80% training** and **20% testing** subsets.
    - Stratified splitting ensured the class imbalance in the target variable was preserved across both sets.
3. **Model Selection**:
    - A **Logistic Regression model** was chosen for its interpretability and simplicity.
    - The model was trained using the preprocessed training data and evaluated on the test data.
4. **Evaluation Metrics**:
    - **Precision**: 0.98 for fraud detection, indicating very few false positives.
    - **Recall**: 0.99, highlighting the model's ability to detect nearly all fraudulent transactions.
    - **F1-Score**: 0.98, balancing precision and recall effectively.
    - **Confusion Matrix**:
        - True Negatives (Non-Fraud Detected): 22,369
        - False Positives (Non-Fraud Misclassified as Fraud): 65
        - False Negatives (Fraud Missed): 21
        - True Positives (Fraud Detected): 2,728
5. **Insights**:
    - The model achieved near-perfect performance on the test set with **99% accuracy**.

    - A small number of false negatives (missed fraud cases) highlights areas for further refinement.

Key Visualizations:

1. **Confusion Matrix**:
    - Visualized true positives, false positives, true negatives, and false negatives to evaluate model performance.
2. **Classification Metrics**:
    - Summarized precision, recall, and F1-score for both fraud and non-fraud classes.

Key Visualizations

1. **Correlation Heatmap**
    - Showed significant negative correlation between `Class` and `Amount` and between `Class` and `Transaction_History`.
2. **Stacked Barplots for Channels and Locations**
    - Highlighted the vulnerability of **online** and **international** transactions.
3. **Age Bracket Fraud Distribution**
    - Illustrated the stark contrast in fraud rates for younger vs. older cardholders.
4. **Boxplots for Transaction Amounts and Histories**
    - Revealed differences in distributions for fraud and non-fraud cases.

Recommendations

1. **Enhanced Monitoring**: Focus on transactions with small amounts at night, particularly in grocery and health categories.
2. **Channel and Location Security**: Strengthen fraud detection mechanisms for online and international transactions.
3. **Demographic-Specific Policies**: Implement stricter fraud prevention measures for younger cardholders (16–25).